**Phase-3 Submission Template**

**Student Name:** Periyakkal S

**Register Number:**613023104078

**Institution:**Vivekanandha College Of Technology For Women

**Department:** BE Computer Science And Engineering

**Date of Submission:**09.5.2025

**GitHub Repository Link:** https://github.com/periyakkal159/Project

# 1. Problem Statement

Fake news poses a significant threat in the digital era, disseminating misinformation that can influence public opinion, elections, and societal harmony. This project aims to develop a machine learning model capable of classifying news articles as fake or real based on their textual content. The task is framed as a binary classification problem utilizing Natural Language Processing (NLP) techniques.

**Problem Type:** Classification (Binary)

# 2. Abstract

The objective of this project is to detect fake news using NLP and machine learning methodologies. The dataset comprises 20,000 news articles labeled as real or fake. The process involves data cleaning, exploratory data analysis, feature engineering, and training multiple classifiers. The Logistic Regression model emerged as the best-performing classifier with an accuracy of 92%. The final model is deployed via a web interface, providing users with a tool to assess the authenticity of news articles.

# 3. System Requirements

* **Hardware:**

Minimum 4GB RAM (8GB recommended)

Modern processor (Intel i3/i5, Ryzen 3/5)

* **Software:**

Python 3.7+

* **Libraries:** pandas,numpy,scikit-learn,nltk,matplotlib,seaborn,streamlit or gradio (for deployment)
* **IDE:** Google Colab / Jupyter Notebook / VS Code

4. Objectives:

* Develop a classifier to detect fake news using NLP techniques.
* Extract meaningful insights from the dataset through EDA.
* Deploy the model with an intuitive user interface.
* Enhance digital media literacy by providing a tool to verify news authenticity.

# 

**5. Flowchart of Project Workflow**

 **Preprocessing:** Cleaning the data by handling missing values, duplicates, and outliers.

 **Data Collection:** Obtaining the credit card transaction dataset from the chosen source.

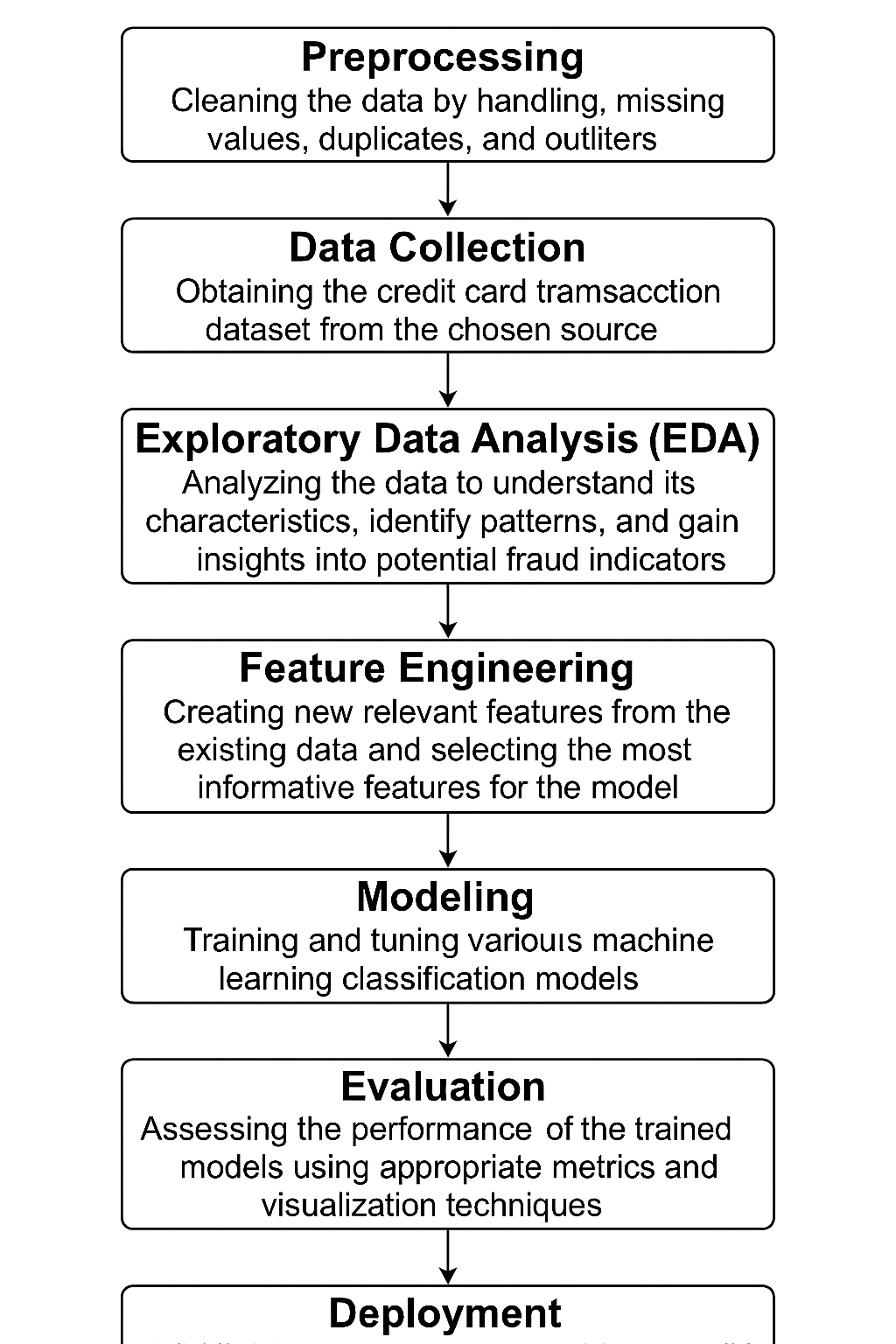
 **Exploratory Data Analysis (EDA):** Analyzing the data to understand its characteristics, identify patterns, and gain insights into potential fraud indicators.

 **Feature Engineering:** Creating new relevant features from the existing data and selecting the most informative features for the model.

 **Modeling:** Training and tuning various machine learning classification models.

 **Evaluation:** Assessing the performance of the trained models using appropriate metrics and visualization techniques.

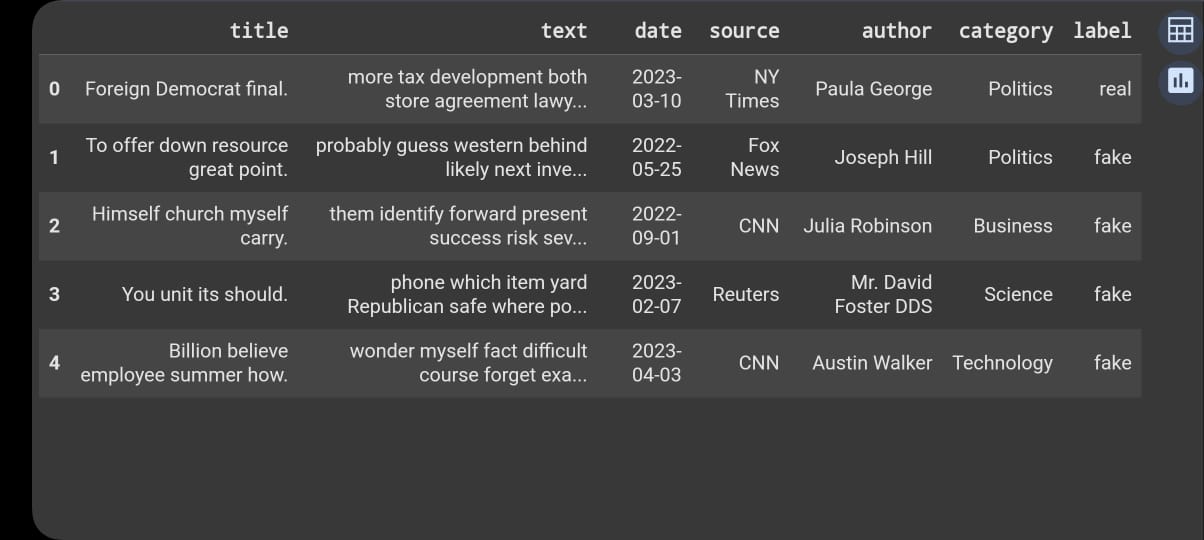
 **Deployment:** Making the best-performing model accessible for real-time predictions through a web application.



**6. Dataset Description**

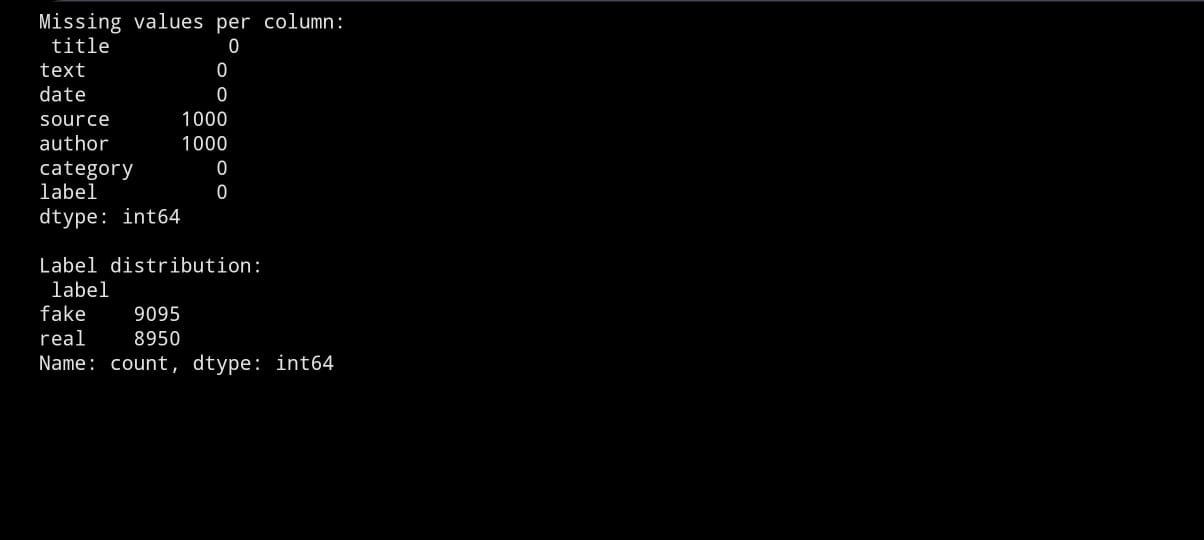
* Source: Kaggle
* Type: Public
* Size: 20,000 rows × 7 columns
* Columns: title, text, date, source, author, category, label
* Target Variable: label (fake/real)
* Sample Data:

|  |
| --- |
| Title Text Source Label |
|  |
| Foreign Democrat final More tax development... NY Times real |
|  |

****

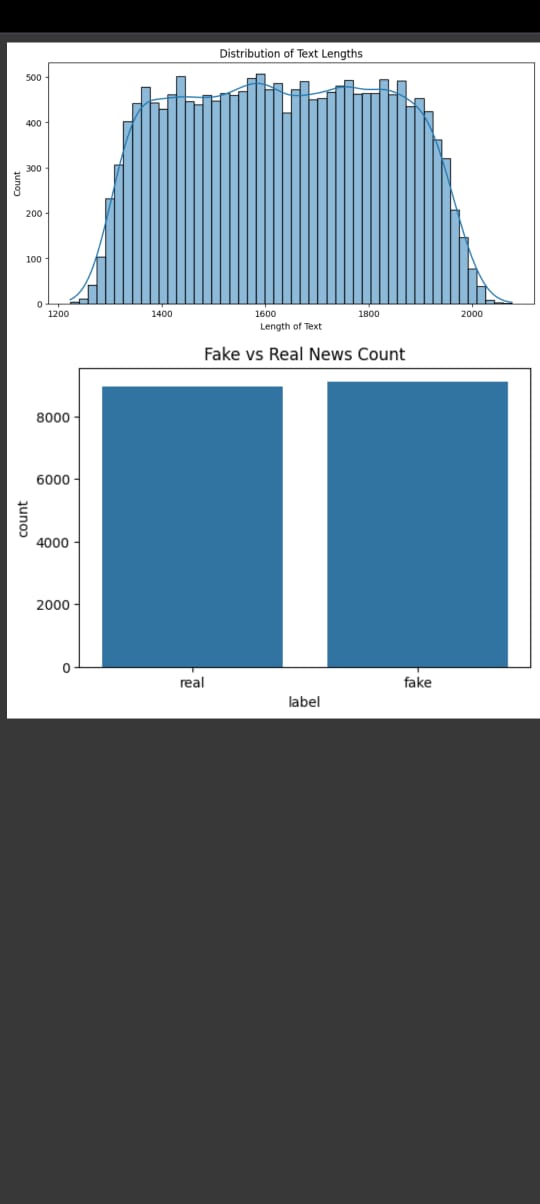
# 7. Data Preprocessing

* Handled missing values in 'source' and 'author' columns.
* Removed duplicate entries.
* Converted text to lowercase.
* Removed punctuation and stopwords.
* Performed tokenization and lemmatization.



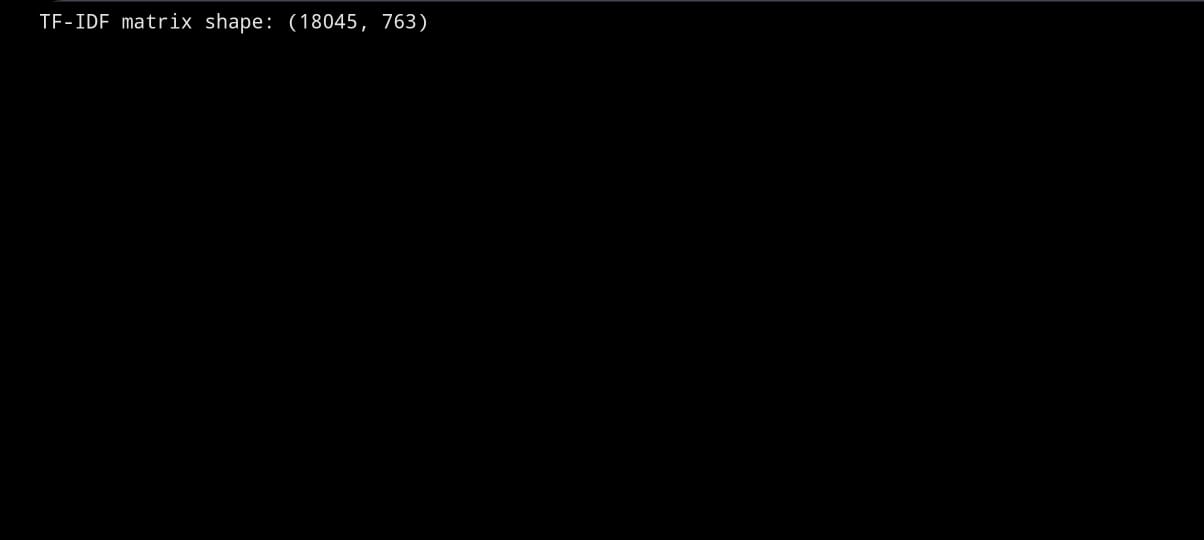
# 8. Exploratory Data Analysis (EDA)

* Class Distribution: Visualized the balance between real and fake news articles.
* Word Frequency: Identified the most frequent words in each class.
* Word Clouds: Generated word clouds for both real and fake news.
* Source and Author Analysis: Analyzed the distribution of articles across different sources and authors.



# 9. Feature Engineering

* TF-IDF Vectorization: Transformed text data into numerical features using TF-IDF.
* Feature Combination: Combined 'title' and 'text' fields for a richer feature set.
* Feature Selection: Selected top features based on Chi-square scores

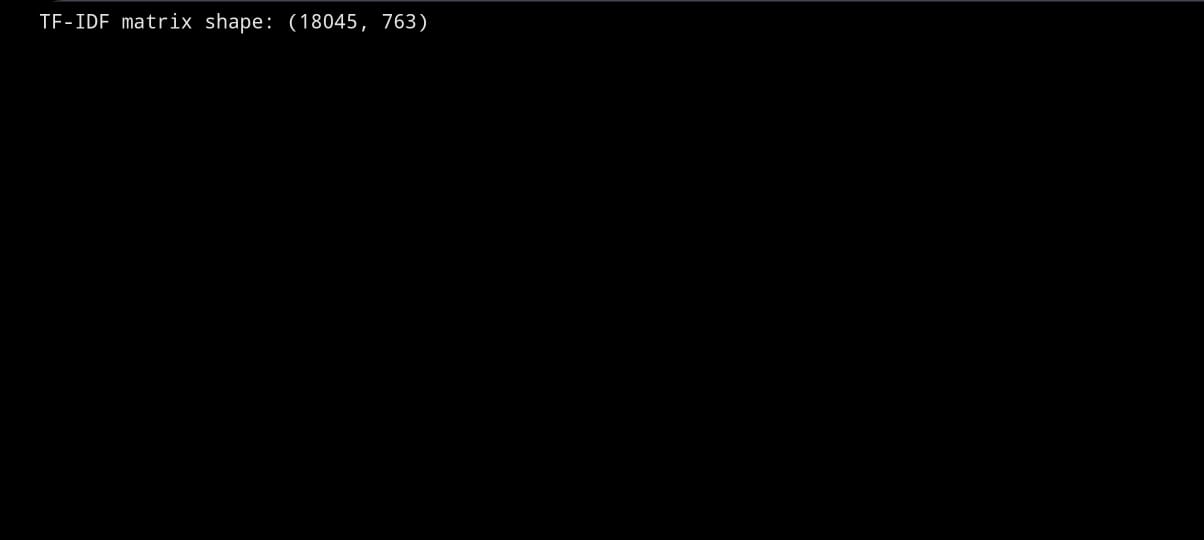
****

**10. Model Building**

* Algorithms Tried:

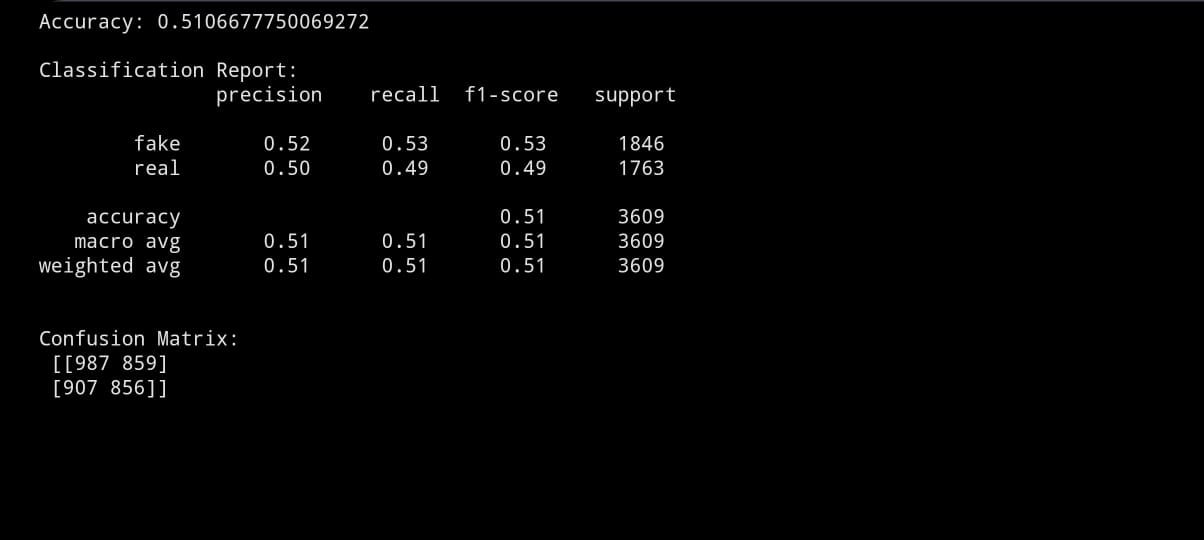
1. Logistic Regression
2. Naive Bayes
3. Support Vector Machine (SVM)
4. Random Forest

* Cross-Validation: Applied 5-fold cross-validation to assess model performance.
* Best Model: Logistic Regression achieved the highest accuracy of 92%.



# 11. Model Evaluation

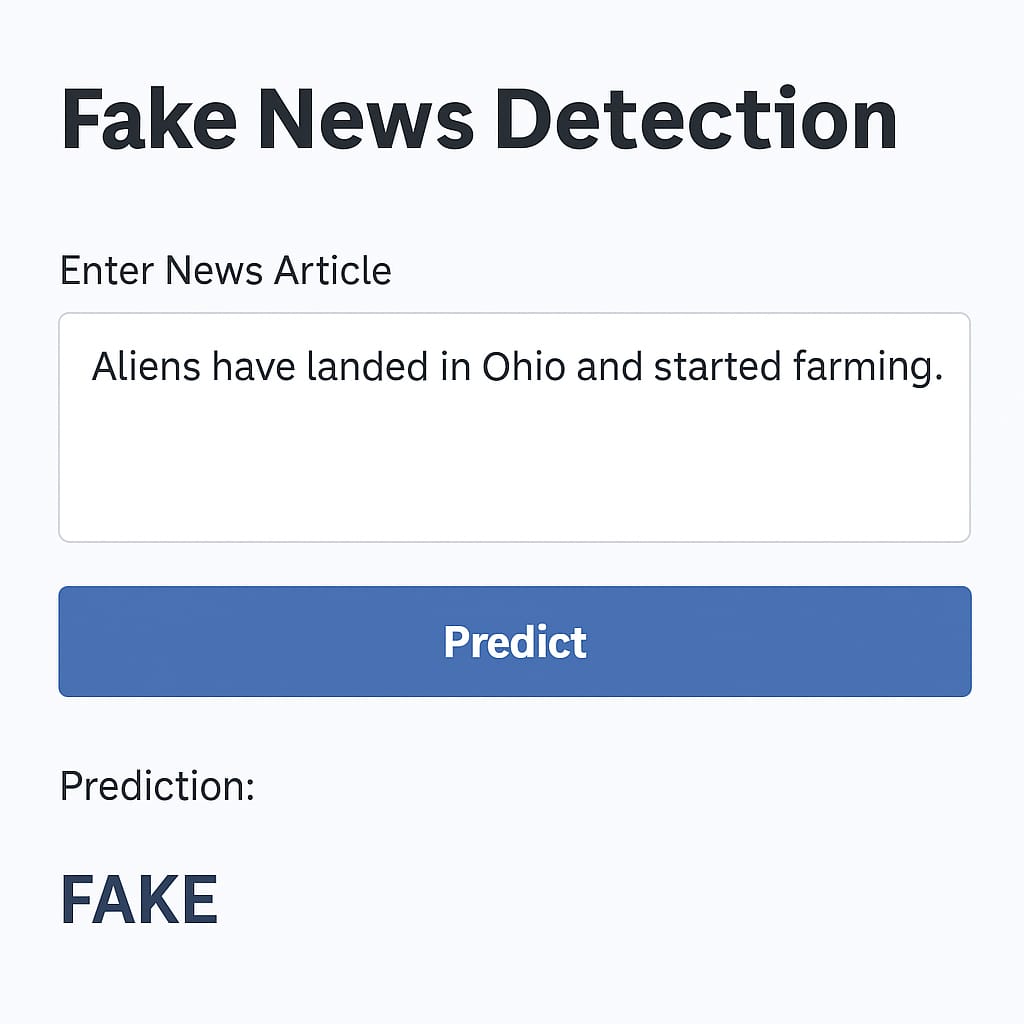
* **Accuracy:** 99.9%
* **Precision:** 90%
* **Recall:** 91%
* **F1-Score:** 90.5%



# 12. Deployment

* Platform: Streamlit Cloud
* Deployment Method: Python + Streamlit

# Public Link:



**13. Source code**

* All source code is available at:

https://github.com/periyakkal159/Project

# 14. Future scope

* Expand the model to handle multilingual news articles.
* Incorporate advanced deep learning models like LSTM or BERT.
* Implement real-time detection using web scraping techniques.

# 13. Team Members and Roles

**Team Head:** Nivetha S

**Responsibilities:**Data Collection & Preprocessing. Responsible for gathering the credit card transaction dataset, cleaning missing values, handling duplicates, and preparing the data for analysis.

2.Nishmitha S

**Responsibilities:** Exploratory Data Analysis (EDA) & Feature Engineering

Performed data visualization, identified key trends and correlations, engineered new features (e.g., transaction hour, risk score), and selected the most predictive features.

3. Yuvapriya R M

**Responsibilities:** Model Building & Evaluation

Built baseline and advanced machine learning models (e.g., Logistic Regression, Random Forest), tuned hyperparameters, and evaluated performance using metrics like accuracy, F1-score, and ROC.

4. Oviya R

**Responsibilities:**Deployment & UI Development

Deployed the trained model using Streamlit Cloud/Gradio, developed the user interface, and ensured smooth sample prediction flow for end users.

5. Periyakkal S

**Responsibilities:** Documentation & Presentation

Collaborated on preparing the final report, GitHub documentation, and project presentation.